Conformalized Survival Distributions
A Generic Post-Process to Increase Calibration

Shi-ang Qi[1], Yakun Yu[1], Russell Greiner[1,2]

[1]University of Alberta, [2]Alberta Machine Intelligence Institute {shiang, yakun2, rgreiner}@ualberta.ca

Objectives

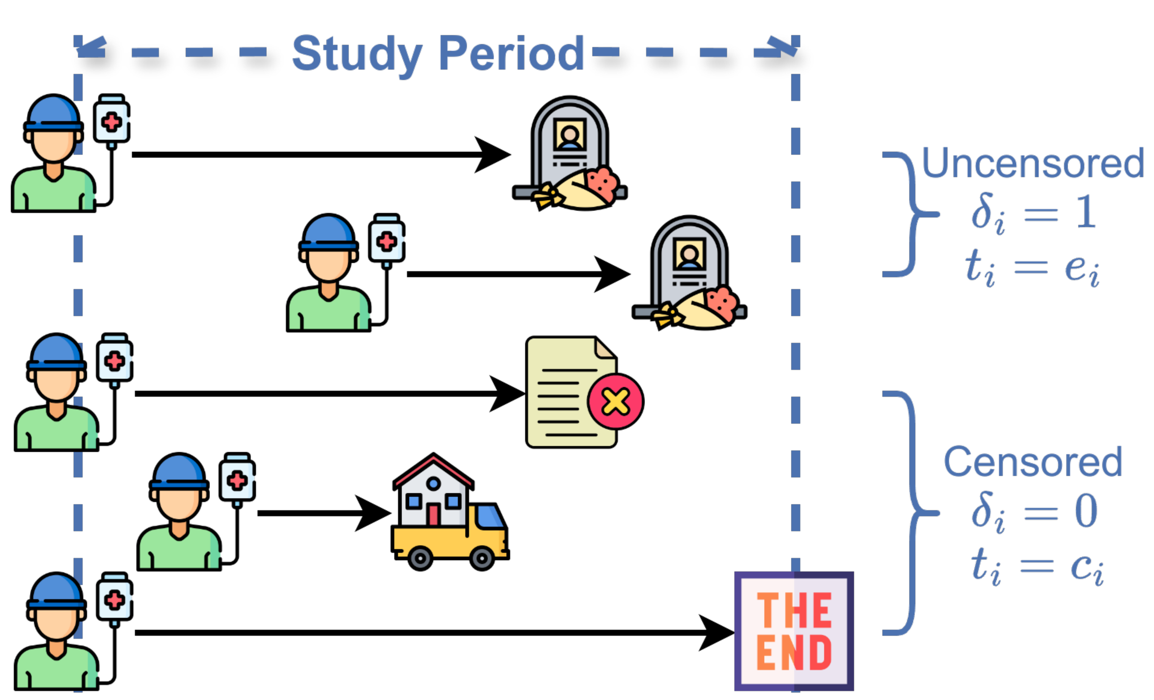
Boost a survival model's calibration ability while maintaining the same discrimination ability.

Survival Analysis

A subject (described $\boldsymbol{x}_i$) is **right-censored** iff it has not experienced an event at the observed time. Each subject is: $[\boldsymbol{x}_i$, observed time t_i, indicator $\delta_i]$, which is based on event time e_i and censor time c_i.

$$t_i \triangleq \min\{e_i, c_i\} \quad \text{and} \quad \delta_i \triangleq \mathbf{1}[e_i \le c_i]$$

Assumptions: (i) **exchangeable** and (ii) **conditional independent censoring**, $e_i \perp c_i \mid \boldsymbol{x}_i$

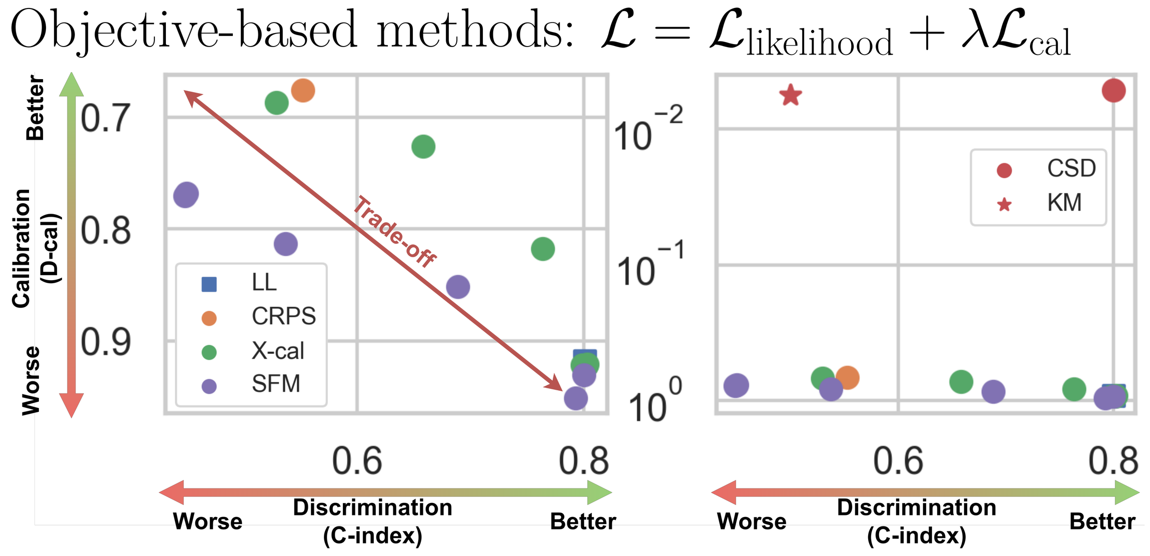


Individual Survival Distribution (ISD) is a probability curve for all future times for a patient:

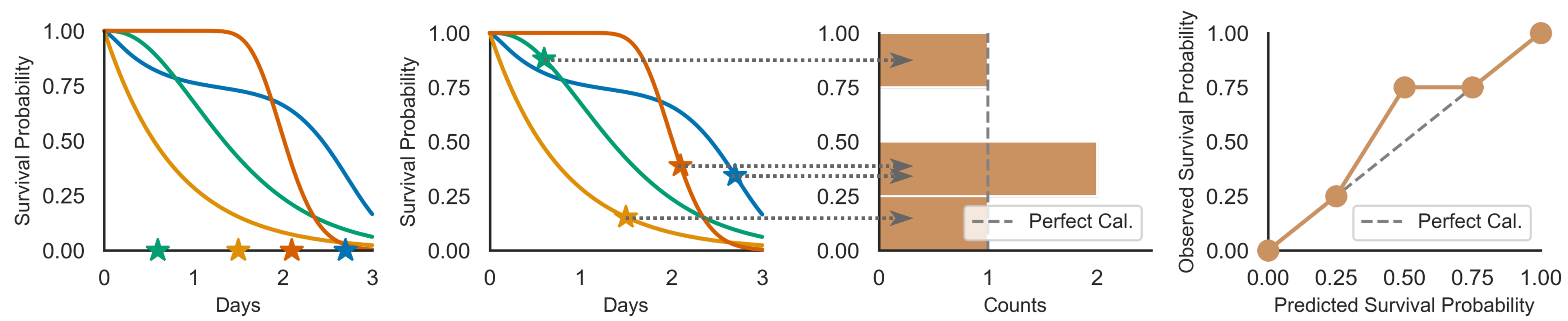
$$S(t \mid \boldsymbol{x}_i) = \Pr(e_i > t \mid \boldsymbol{x}_i).$$

Discrimination vs Calibration

Discrimination: ability to accurately rank subjects.
Calibration: predicted probs. match the obs.
Objective-based methods: $\mathcal{L} = \mathcal{L}_{\text{likelihood}} + \lambda \mathcal{L}_{\text{cal}}$

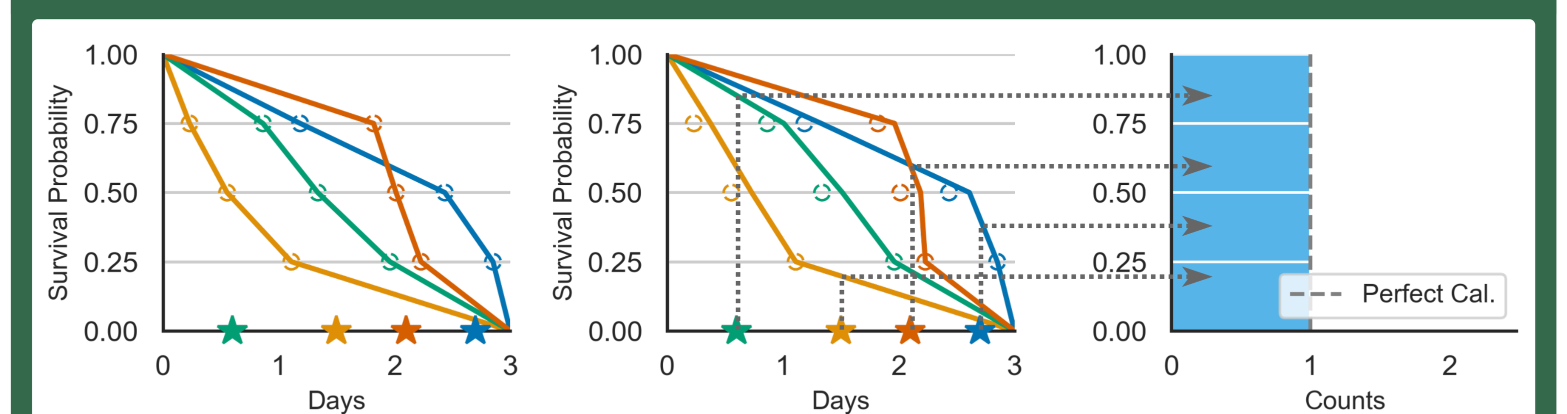


Calibration in Survival Analysis



⇒ **Distribution calibration (D-cal)**[1]: the predicted survival probability at true event time, $\{\hat{S}(e_i \mid \boldsymbol{x}_i)\}_i$, should follow $\mathcal{U}[0,1]$ (*inverse transform theorem*). For a censored subject, it follows $\mathcal{U}[0, \hat{S}(c_i \mid \boldsymbol{x}_i)]$.
⇒ **KM calibration (KM-cal)**: the average predicted ISD should align with the empirical survival distribution for the dataset (Kaplan-Meier curve). See the paper for a visual example.

Conformalized Survival Distribution



Main steps of CSD (for uncensored subjects):

❶Split data to a training set $\mathcal{D}_{\text{train}}$ and a conformal set $\mathcal{D}_{\text{con}}$.
❷Learn a model $\mathcal{M}$ from $\mathcal{D}_{\text{train}}$ and predict ISDs for $\mathcal{D}_{\text{con}}$.
❸Discretize the ISD predictions at predefined percentile levels ρ.

$$\hat{q}_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i) = \inf\{t : \hat{S}_{\mathcal{M}}(t \mid \boldsymbol{x}_i) \le \rho\} = \hat{S}^{-1}_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i),$$

❹Apply conformal quantile regression [2] at each percentile level.
❶$s_{i,\mathcal{M}}(\rho) = \hat{q}_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i) - t_i, \quad \mathcal{S}_{\mathcal{M}}(\rho) = \{s_{j,\mathcal{M}}(\rho)\}_{j=1}^{|\mathcal{D}_{\text{con}}|},$
❷$\hat{q}'_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i) = \hat{q}_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i) - \text{Quantile}[\rho; \mathcal{S}_{\mathcal{M}}(\rho)],$
❺Apply rearranging methods to $\hat{q}'_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i)$ and transform to ISDs.

Theoretical Guarantees

Discrimination Applying the CSD adjustment does not affect the Harrell's C-index of the model.
D-cal Under the assumptions (i) & (ii), CSD exhibits exact D-cal.

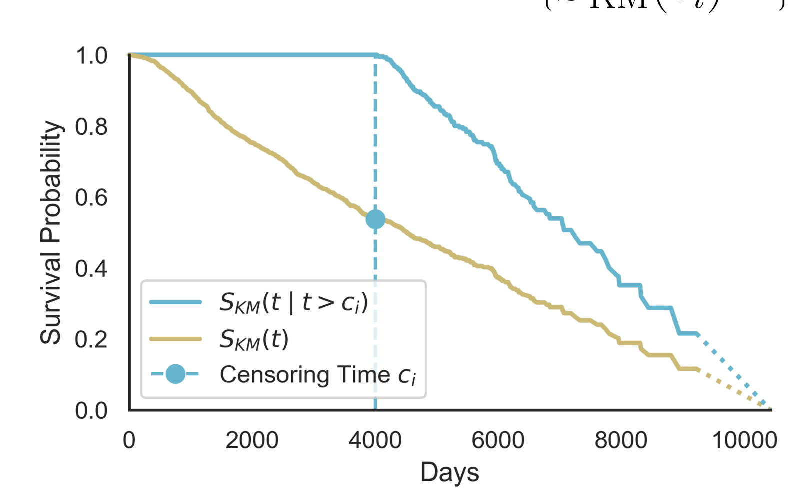
$$\forall \rho, \quad \rho \le \Pr(t_i \in [\hat{q}'_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i), \infty] \mid \boldsymbol{x}_i) \le \rho + 1/(|\mathcal{D}_{\text{con}}| + 1).$$

KM-cal Under the assumptions (i) & (ii), CSD asymptotically exhibits exact integrated calibration at all time points.

KM-Sampling

Problem: We do not know t_i for censored subjects (step 4).
Solution: Approximate the uncertainty of t_i using a "best-guess distribution", and sampling surrogate times from it:

$$S_{\text{KM}}(t \mid t > c_i) = \min\left\{\frac{S_{\text{KM}}(t)}{S_{\text{KM}}(c_i)}, 1\right\}.$$



Repeat every subject R times:

$$s^r_{i,\mathcal{M}}(\rho) = \hat{q}_{\mathcal{M}}(\rho \mid \boldsymbol{x}_i) - t^r_i,$$

For a censored subject:

$$t^r_i \sim S_{\text{KM}}(t \mid t > c_i).$$

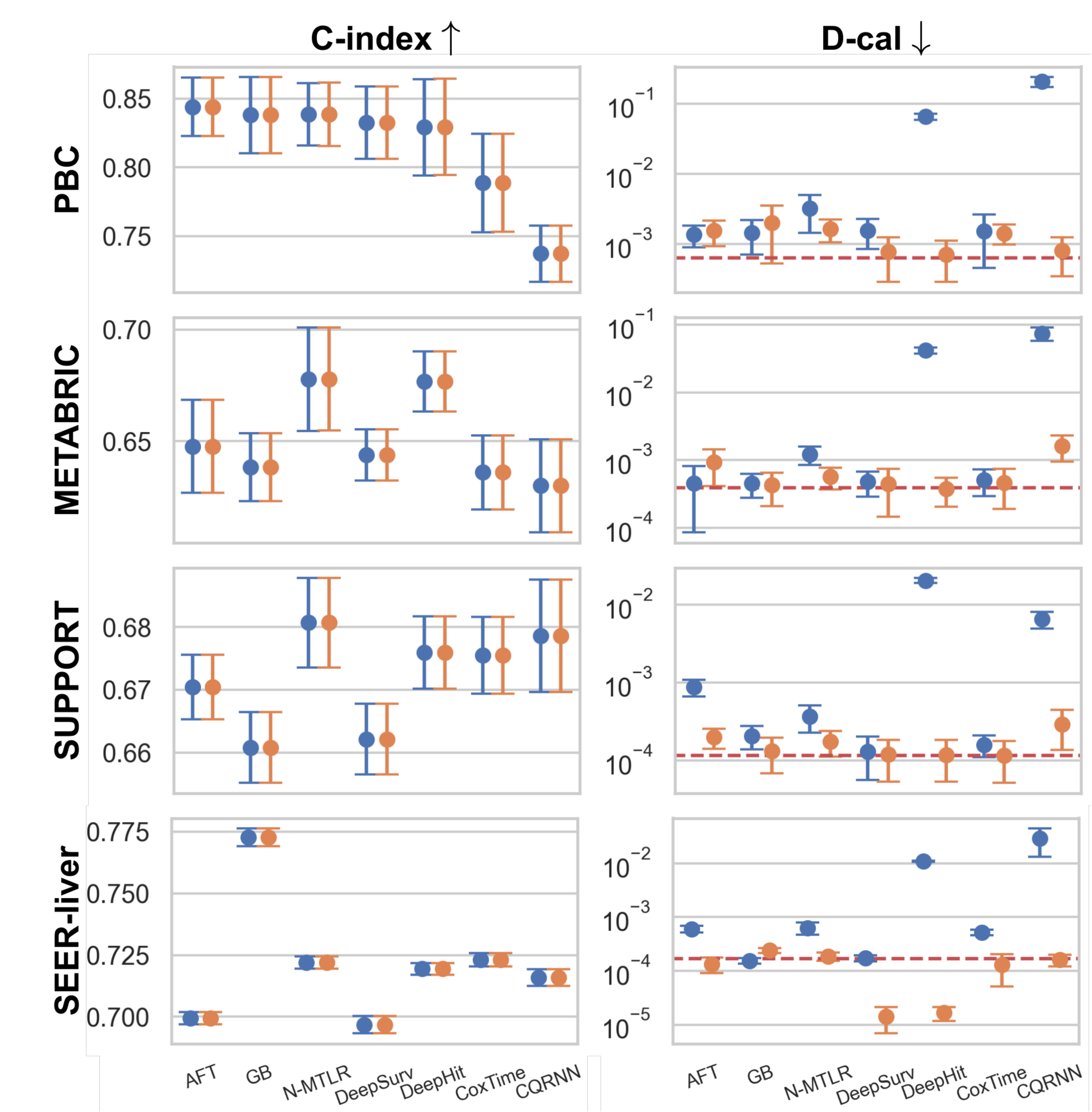
For an uncensored subject:

$$t^r_i \leftarrow e_i$$

Why choose KM? It is guaranteed to be asymptotically calibrated (both D-cal and KM-cal).

Empirical Results

Blue: baseline, **Orange:** baseline + CSD
Red dash line: empirical lower bound for calib.



Using 11 real datasets and 7 baselines, we have 76 comparisons (AFT does not converge for 1 case).

	C-index	D-cal	KM-cal	IBS	MAE-PO
Non-CSD	3(0)	8(1)	20(7)	12(0)	30(0)
CSD	13(0)	**68(35)**	**56(30)**	**61(14)**	**45(4)**
ties	**60**	0	0	3	1

† Number of wins (Number of significant wins with $p < 0.05$).

Findings from ablation studies:

- KM sampling outperforms other naive methods.
- For small size data, we should reuse $\mathcal{D}_{\text{train}}$ in the conformal step to maintain discriminative power.
- Different values of ρ have minimal impacts.

References

[1] Humza Haider et al. Effective ways to build and evaluate individual survival distributions. JMLR 2020
[2] Yaniv Romano et al. Conformalized Quantile Regression. NeurIPS 2019